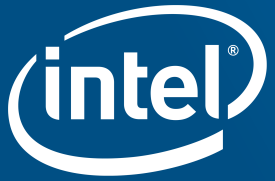


BEYOND SPEEDS AND FEEDS - WHAT NEW CAPABILITIES ARE NEEDED FOR EXASCALE INTERCONNECTS?

Jeff Hammond

Parallel Computing Lab, Intel Corporation

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS”. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Optimization Notice

Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Extreme Scalability Group Disclaimer

I work in Intel Labs and therefore don't know anything about Intel products.

I am not an official spokesman for Intel.

I do not speak for my collaborators, whether they be inside or outside Intel.

You may or may not be able to reproduce any performance numbers I report.

Hanlon's Razor (blame stupidity, not malice).

Hardware

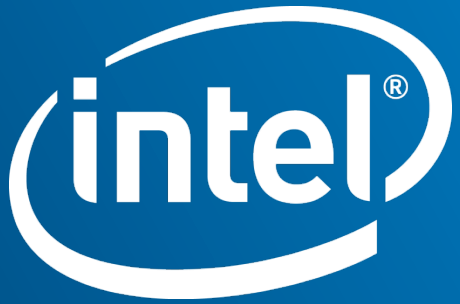
- Endpoint resource partitioning (primarily for multithreading).
- Standard atomics, including 64b float math. Whatever MPI3+ needs.
- Pre-posted receives should make asynchronous progress.
- Support fabric isolation and/or QoS.
- Performance monitoring tools everywhere.
- Power-related features.
- Efficient reliability. Handle node faults effectively. Prevent Byzantine failures.

Software

- MPI-4 is the only necessary exascale ~~programming model~~ runtime system.
- Unrestricted multithreading support should be maximally efficient.
  - Lockless. MPI endpoints? Per-communicator queues?
- Send-Recv is a necessary evil. Emphasize scalable/efficient constructs like neighborhood collectives and RMA.
- Nonblocking must mean asynchronous.
- Fault-tolerance. Checkpoint-restart inadequate for many workloads.
- Efficient interaction with the processor without heavy polling (for power).

Software

- Composability
  - MPI, PGAS, IO/Analysis, Profile/Debug as separate clients of network API.
  - Bootstrap MPI on top of cloud frameworks (MPI sessions?).


intel®